Assessment of physicians' professional performance

ENGLISH SUMMARY

The assessment of practicing physicians is common around the world, with the aim to help physicians improve their performance and -ultimately- to improve health care. It is generally acknowledged that the assessment of and feedback on physicians' performance is critical to the development (and maintenance) of their expertise. For the assessment methods to be meaningful for feedback, and to reach justified high-stake decisions on physician performance, they should provide valid results. Validity is the sine qua non of assessment results; without validity, assessment results have little or no meaning. As introduced in Chapter 1, an often-used method to assess the performance of practicing physicians are questionnaire-based tools (QBT), including multisource feedback (MSF). Not surprisingly, research on MSF focused on its validity and mostly concluded that this type of tool have validity. However, essential nuances were lacking from results and conclusions of this research, as stated in chapter 1. Validity is concerned with justifying specific uses of assessment results, and not whether the assessment tool is valid. Validity is concerned with whether it is justified to use the assessment results, for example, for formative feedback or for summative decisions. This requires prioritization of specific validity evidence, instead of gathering all sorts of evidence. Furthermore, various notions that exist upon the underlying ontological definition of physicians' professional performance requires a neutral validity framework. A neutral validity framework is not restricted bounded to a particular epistemological stance and accepts trustworthy evidence of different epistemological stances, to strengthen the validity argument.

The primary aim of this thesis was to understand how valid the results of questionnaire-based assessment methods are for formative and summative reasons for practicing physicians, using a neutral validity framework. To reach this aim, the following research question was addressed: "What evidence is there to be collected, to support or refute the validity argument of questionnaire-based assessments of physicians' professional performance, for formative and summative purposes?". For this end, this thesis treats validity as an argument. With this argument-based approach, an argument for validity must be made and the different aspects of the validity argument should be considered.

In **chapter 2**, all aspects of the validity argument have been considered in a systematic review of the literature on questionnaire-based tools for assessing practicing physicians. The four aspects to be considered for the validity argument are scoring, generalization, extrapolation and implications, and all four taken together should create a coherent chain. The scoring aspect of the argument requires evidence that the 'scoring' of the observations is appropriately done, thus whether the assessment items, scores and assessors are appropriate for the assessment. Generalization takes the scoring aspect further and requires evidence of whether the assessment results would be reproducible

in a different assessment setting. Extrapolation is concerned with finding evidence of validity outside the assessment setting, thus whether the results produced from the assessment would extrapolate to 'real world' performance. Lastly, the implications part of the validity argument implies that the resulting consequences of the assessment are reached, and no unintended consequences are overlooked. With a systematic search of the literature on QBT, 15 tools were found that were described in 46 research articles. Besides these tools, that were specifically aimed at evaluating physicians' performance in clinical practice, we also searched for tools aimed at assessing physicians' teaching and research performance. Thirty-eight tools were available from the literature to assess physicians' clinical teaching performance. However, no tools were available to assess physicians' research performance. With this review we gathered all the available evidence on the four validity aspects: scoring, generalization, extrapolation, implications. We concluded that not every aspect had received sufficient attention in the quest for validity, especially when considering the summative use of these tools. In essence, the evidence of the scoring aspect of questionnaire-based tools seems troublesome when regarding that 'scorers' or the assessors of physicians' professional performance are 'subjective'. Furthermore, there was a lack of evidence surrounding the implications aspect of the argument. Whether physicians improved after the assessment has not been investigated in-depth; the focus was mostly on self-reported evidence. With this review, the weakest links in the argument were identified and provided focus to our subsequent research.

Chapter 3 reports on the validity evidence for the questionnaire-based tool 'Inviting Coworkers to Evaluate Physicians Tool' (INCEPT), a tool intended to assess physicians and provide them with formative feedback on their performance. To further examine the strength of the validity argument for questionnaire-based tools, an approach was needed that encompasses that different assessors capture different views of physicians' professional performance. In this study, 218 physicians were assessed by 597 peers, 344 residents and 822 coworkers; they received 3223 evaluations in total. By conducting exploratory and confirmatory factor analyses, we investigated how the three different assessor groups perceived physicians' professional performance and analyzed how these three groups differ in their clustering of performance domains. The results of the factor analyses showed an acceptable to good fit with three factors for all three assessor groups: assessors perceived physicians' performance to include showing a 'professional attitude', showing 'patient-centeredness' and possessing 'organization and (self) management' skills. The clustering of these performance domains differed slightly per assessor group, thus showing that the assessor groups perceive physicians' professional performance differently. The 3-factor solution was further supported by the item-total correlations >0.50, indicating that each item contributes to the measurement of professional performance, and inter-scale correlations <0.79 indicating that the INCEPT domains overlapped by less than 60%. Evidence of extrapolation was further

established by significant positive associations between numerical and narrative feedback of assessors. This association indicates that the more positive comments were given to a physician, the higher this physician's total INCEPT score was. Likewise, the more suggestions for improvement were given, the lower the physician's INCEPT score. The results of generalizability analyses showed that a minimum of three peers, two residents and three coworkers are needed to assess the overall professional performance reliably.

The next step in investigating the validity argument of questionnaire-based tools was to examine a gap in the extrapolation aspect. A lack of research on the associations between physicians' 'subjective' MSF scores and 'objective' clinical outcomes fueled the study reported in **chapter 4**. With this study, we examined whether anesthesiologists who perform well on clinical outcome measures would also receive higher ratings from their assessors with MSF. In 2014, 28 anesthesiologists from one academic hospital, who performed 8030 anesthetic procedures, were evaluated with MSF by 56 residents, 38 peers, 69 consultants from other specialties, and 144 coworkers. With MSF data resulting from the 'INCEPT', we determined associations between anesthesiologists' mean scores on the three performance domains - professional attitude, patient-centeredness, organization and (self)management - and several 'Quality of Care' (QoC) measures. These measures were predefined by literature, research and protocols. They included anesthesiologists' average performance on three outcome and two process measures, namely anesthesiologists' (1) intraoperative pain management, (2) prevention of postoperative nausea and vomiting, (3) intraoperative temperature monitoring, (4) normothermia management and (5) neuromuscular function monitoring. With multilevel regression analyses we found several significant associations between the ratings given and anesthesiologists' QoC measures. We found that anesthesiologists who performed well on intraoperative temperature monitoring and prevention of postoperative nausea and vomiting, received higher patient-centeredness ratings from all assessor groups. Anesthesiologists who better maintained patients' normothermia received higher professional attitude ratings by residents but received lower ratings from coworkers. Residents gave higher organization and (self)management ratings to anesthesiologists who monitored patients' intraoperative temperature better, whereas other specialty-consultants gave lower ratings to these anesthesiologists. These findings show that the associations between subjective MSF ratings and objective clinical outcome measures are not that straightforward. Although every assessor group agrees that the anesthesiologists who monitor intraoperative temperature and prevent postoperative nausea and vomiting, the higher their patient-centeredness score should be, for the other professional domains the associations between the measures are less straightforward.

The final step in the validity argument scrutinization was to explore possible evidence of the implications component: what are the consequences for physicians' subsequent professional performance after physicians receive MSF on their performance? With MSF, it is believed that physicians can improve their performance after receiving the feedback as it reveals shortcomings in current performance, while at the same time performance can be praised. The observational study described in **chapter 5** investigates evidence of this last component by looking at 103 physicians' MSF scores over time. These physicians were evaluated twice with MSF, by 242 residents, 684 peers and 999 coworkers, while completing a self-evaluation as well. In this study, we specifically looked at the possible consequences of divergent feedback, namely when physicians rated themselves higher in the MSF than their assessors. Within MSF evaluations, physicians can be confronted with feedback that is incongruent with their own performance beliefs. This incongruence can either be positive or negative, meaning that physicians either underrated or overrated their own performance, respectively. Especially negative discrepancies between self-assessment scores and assessors scores are interesting to consider when looking at the consequences of MSF, since they can either stimulate behavioral change or be destructive for future performance. On the one hand, negative discrepancies between physicians' self-assessment scores and assessors' assessment scores are beneficial for physicians as they reveal current, unknown, performance gaps. On the other hand, when confronted with negative discrepancies, physicians may also experience emotional distress that might be unfavorable for physicians' subsequent performance changes. Up till now, little was known about the influence of these negative discrepancies on physicians' professional performance. Using mixed-effects models, we quantified the associations between negative discrepancies and the change in subsequent MSF scores for physicians, in three performance domains: 'professional attitude', 'organization and (self)management' and 'patient-centeredness'. The outcome of interest was physicians' average domain score changes, thus the change in scores between the first and second MSF evaluation. Considering the differences between assessor groups, we differentiated between the scores that residents, peers and coworkers gave to the same physician. The predictor variable, negative discrepancy score, was calculated as how many times physicians overrated themselves on feedback items, compared to the average item score given by residents, peers and coworkers. This variable ranged from zero to 18, indicating that when physicians never overrated themselves a negative discrepancy score of zero was given, as opposed to when physicians overrated themselves on every item resulting in a score of 18. After controlling for physicians' and evaluations' characteristics, the results show that negative discrepancies are negatively associated with score changes in all three professional performance domains. This means that when physicians are confronted with negative discrepancies, the extent of physicians' performance improvement declines, and at one point, even performance decline occurs. Physicians' confidence in own performance might explain this phenomenon, as too much self-

confidence has been shown to cause more frequent dismissal of feedback. This result calls for extra attention for physicians who overrated themselves, when they receive their feedback report.

In **chapter 6** the results of the previous studies were summarized, synthesized and considered in light of two epistemological stances to enhance the depth of the complex topic of assessment of physicians' professional performance. This chapter provides the answer to our research question: "What evidence is there to be collected, to support or refute the validity argument of questionnaire-based assessments of physicians' professional performance, for formative and summative purposes?". The answer to this question is not straightforward nor easily summarized. The different epistemological stances existing within the framework of physicians' professional performance assessment call for different considerations with respect to the answer to the research question. Although both research paradigms focus differently on the validity evidence, from both stances it can be concluded that the validity argument of using questionnaire-based tools, including multisource feedback, for summative reasons is not strong enough yet. We proposed an alternative assessment design to advance the use of questionnaire-based tools for formative and summative purposes: the model of programmatic assessment. Programmatic assessment asks for various assessment components that are thoughtfully combined and constructed as a program of assessment, intended to capture the complete and complex performance of the physician. We provided recommendations for using this model of assessment in practice and a plan for future research on this type of assessment. Furthermore, we stated that the answer to our research question and the generalization of the results should be viewed while taking the limitations of the present studies into account. This chapter ends with a saying: "Great minds think alike - but fools rarely differ". Although this saying is meant to indicate that when two people have the same idea, they could be either brilliant or foolish, I like to say that indeed great minds may think alike, but only fools would rarely differ in their perspective.

DUTCH SUMMARY

Dit proefschrift is geschreven naar aanleiding van de publieke belangstelling voor het professionele functioneren van artsen. Daarbij richt dit onderzoek zich met name op de validiteit van de beoordeling van het professioneel functioneren van praktiserende artsen. De beoordeling van het professioneel functioneren van praktiserende artsen is van groot belang voor zowel artsen zelf als hun patiënten. Het kan artsen, daar waar nodig is, ondersteuning bieden om hun functioneren te verbeteren, met als uiteindelijk doel de gezondheidszorg te verbeteren.

Feedback op het functioneren van artsen is essentieel voor de ontwikkeling (en het onderhoud) van hun expertise. Echter om zinvolle feedback te geven aan artsen, moet deze feedback wel valide zijn. Hetzelfde geldt voor het maken van belangrijke beslissingen over artsen hun functioneren (zoals herregistratie voor medisch specialisten); ook deze moeten valide zijn. Validiteit is de sine qua non van beoordelingen, of liever gezegd de resultaten resulterende uit beoordelingen. Zonder validiteit hebben beoordelingsresultaten weinig of überhaupt geen betekenis. Zoals geïntroduceerd in **hoofdstuk 1**, worden vragenlijst methoden, waaronder 360° feedback, oftewel multisource feedback (MSF), veel gebruikt om het functioneren van artsen te evalueren en te beoordelen. Met MSF kunnen artsen hun functioneren laten evalueren en beoordelen door verschillende groepen –collega's, patiënten, studenten- een vragenlijst te laten invullen. Deze beoordelaars die de arts in de praktijk kunnen observeren, geven dan op basis van een vragenlijst, scores en geschreven feedback aan artsen. Het is wellicht niet verrassend dat onderzoek naar MSF zich vooral concentreerde op de validiteit ervan. Voorgaand onderzoek concludeerde dat dit soort methodes, vragenlijsten en MSF, validiteit bezitten. Echter, er ontbraken belangrijke nuances in de onderzoeksresultaten en daaruit getrokken conclusies. Zo was het niet duidelijk voor welk doel het instrument precies valide was. Is het gebruik van vragenlijst methodes valide om te gebruiken voor het geven van feedback, en voor het maken van belangrijke beslissingen over artsen hun functioneren? Validiteit, of valideren, is het proces van rechtvaardigen van het specifieke gebruik van beoordelingsresultaten, en betekent niet dat de specifieke beoordelingsmethode valide is. Bij validiteit gaat het erom of het terecht is om de beoordelingsresultaten te gebruiken voor verschillende doeleinden. De doelen voor het gebruik van vragenlijsten om het functioneren van artsen te beoordelen verschillen ook. Het doel van vragenlijsten om artsen hun functioneren te evalueren is om feedback te geven, terwijl bij beoordelen het uiteindelijk doel is om belangrijke beslissingen te maken. Het ene doel vraagt ander bewijs dan het andere doel. Deze verschillende doeleinden vereist het prioriteren van bepaald soort validiteitsbewijs, in plaats van het lukraak verzamelen van allerlei bewijsmateriaal. Bovendien bestaat er onenigheid over de onderliggende definitie van het professionele functioneren van artsen. Zo ziet één perspectief, het post-

positivistische perspectief, het functioneren van artsen als meetbaar waarbij er een ware score te meten is. Terwijl het socio-constructivistisch perspectief het functioneren van artsen niet als één ware score ziet, maar dat het functioneren van artsen interpersoonlijk en niet direct meetbaar is. Deze verschillende perspectieven op het functioneren van artsen vragen om een neutraal validiteitskader in het onderzoek naar validiteit. Een neutraal validiteitskader is namelijk niet gebonden aan één bepaald wetenschapskader en accepteert betrouwbaar bewijs vanuit verschillende perspectieven.

Het primaire doel van dit proefschrift was om te onderzoeken, met een neutraal validiteitskader, hoe valide de resultaten van op vragenlijsten gebaseerde beoordelingsmethoden zijn voor het evalueren en beoordelen van praktiserende artsen. Om dit doel te bereiken, werd de volgende onderzoeksvraag gesteld: "Welk bewijs moet er worden verzameld, ter ondersteuning of weerlegging van het validiteits-argument voor het gebruik van vragenlijsten om artsen hun functioneren te evalueren en te beoordelen?" Daartoe werd validiteit gezien als het maken van een argument, waarbij verschillende onderdelen van dat argument allen in overweging genomen moeten worden. Door alle onderdelen van dit validiteitsargument van voldoende en kwalitatief sterk bewijs te voorzien, kan er een sterk argument gemaakt worden voor de validiteit van het gebruiken van een beoordelingsmethode.

In **hoofdstuk 2** is er onderzoek gedaan naar het validiteitsbewijs van alle bestaande vragenlijsten in de literatuur. Specifiek is hierbij gekeken of er genoeg bewijs was voor de vier verschillende onderdelen van het validiteitsargument: scoren, generaliseren, extrapoleren en implicaties. Het onderdeel 'scoren' vraagt bewijs dat het 'scoren' van de observaties goed is toegepast. Oftewel, of de vragen/items, scores en beoordelaars geschikt zijn voor het scoren van het professioneel functioneren van de praktiserende arts. Het volgende onderdeel in het argument gaat over 'generaliseren'; kunnen we de resultaten die zijn behaald in de ene evaluatie/beoordeling-setting, reproduceren in een andere evaluatie/beoordeling-setting. Het gaat om de vraag of de arts met de gekozen vragen/items, scores en beoordelaars dezelfde resultaten zou verkrijgen als er andere vragen/items, scores en beoordelaars zouden zijn gebruikt. Voor bewijs met betrekking tot het extrapoleren van de resultaten kijken we naar het daadwerkelijke gedrag in de praktijk, in plaats van alleen naar het functioneren zoals gezien in de evaluatie/beoordeling-setting. Het gaat erom of de arts, die geobserveerd werd in een beoordeling-setting ook hetzelfde zou functioneren als deze niet geobserveerd werd. Het laatste onderdeel van het argument focust op de implicaties van de behaalde resultaten, en wat voor beslissingen op basis van deze resultaten worden genomen. Zijn de implicaties, resulterende uit deze beslissingen, wel rechtvaardig? Verbeteren artsen hun functioneren na het verkrijgen van feedback? Of zijn er onbedoelde consequenties verbonden aan de genomen beslissingen?

Met het gebruik van een systematisch literatuur onderzoek naar vragenlijsten is er getracht bewijs te verzamelen voor de vier onderdelen van het validiteitsargument. Met dit onderzoek zijn 15 vragenlijsten gevonden, beschreven in 46 artikelen. Naast deze vragenlijsten, die ontworpen waren om het functioneren van artsen in hun rol als zorgverlener te evalueren en te beoordelen, zijn we ook op zoek gegaan naar vragenlijsten voor het beoordelen van artsen in hun rol als opleider en als onderzoeker. Er zijn 38 vragenlijsten gevonden om artsen in hun rol als opleider te evalueren en te beoordelen, echter voor artsen in de rol van onderzoeker zijn geen vragenlijsten gevonden. Alle vragenlijsten en de bijbehorende validiteitsbewijzen zijn onder de loep genomen, waarbij er geconcludeerd moest worden dat er nog onvoldoende bewijs is om het gebruik van vragenlijsten bij de beoordelingen van artsen te rechtvaardigen, vooral wat betreft het gebruik van vragenlijsten om belangrijke beslissingen over artsen hun functioneren te maken. Er blijkt dat voor het onderdeel 'scoren' nog onduidelijkheid bestaat over de geschiktheid van de beoordelaars: het lijkt erop dat deze te 'subjectief' zijn om geschikte beoordelaars te zijn voor praktiserende artsen. Ook voor het onderdeel 'implicaties' schort er nog het één en ander: er is weinig bewijs of artsen daadwerkelijk hun functioneren verbeteren na het krijgen van feedback. Ook bleek een belangrijk aspect van het onderdeel 'extrapoleren' niet voldoende onderzocht, namelijk hoe de beoordelingen van artsen, gebaseerd op vragenlijsten, relateren aan hun daadwerkelijke klinische functioneren. Met dit onderzoek hebben we de zwakste onderdelen van het validiteitsargument blootgelegd, en zo ook richting gegeven aan ons verdere onderzoek.

Hoofdstuk 3 gaat in op het verzamelen van validiteitsbewijs voor het gebruik van een specifieke multisource feedback tool, gericht op het evalueren van artsen om zo feedback te geven op hun functioneren. Deze tool, de '*INviting Coworkers to Evaluate Physicians Tool*', ofwel de 'INCEPT', is zo ontworpen dat drie verschillende soorten beoordelaars één en dezelfde vragenlijst gebruiken. Zo gebruikten collega medisch specialisten, artsen in opleiding (AIOS), en andere medewerkers (de drie type beoordelaars) één en dezelfde vragenlijst. De INCEPT was enigszins praktisch ingesteld, omdat artsen zo gemakkelijker hun beoordeling op basis van deze ene vragenlijst konden doornemen, in plaats van drie verschillende vragenlijsten. De analyses naar de validiteit zijn echter wel per type beoordelaar verricht. Op basis van resultaten uit beoordelaars-expertise onderzoek bleek het noodzakelijk om de drie verschillende soorten beoordelaars hun eigen perspectief op het functioneren van artsen te laten houden. In deze studie waren 218 artsen vanuit verschillende ziekenhuizen en specialismen, beoordeeld door 597 collega medisch specialisten, 344 AIOS en 822 medewerkers, die in totaal 3223 beoordelingen hebben gegeven. Door middel van hiervoor geschikte statistische methoden, zoals factoranalyses, is onderzocht hoe de vragen van de vragenlijst bij elkaar clusteren in verschillende domeinen, rekening houdend met de drie verschillende type beoordelaars. Voor alle drie de typen

beoordelaars werd een acceptabele tot goede fit gevonden voor drie verschillende domeinen. De vragenlijst is onder te verdelen in drie domeinen, waarbij het functioneren van artsen gezien wordt als 'patiëntgerichtheid', 'professionele houding' en '(zelf)management en organisatorische vaardigheden'. De vragen die bij deze verschillende domeinen behoren, verschilden lichtelijk per type beoordelaar. Het bewijs voor deze drie domeinen werd verder ondersteund door de gevonden item-totaalcorrelaties, die allen onder de 0,50 waren. Dit geeft aan dat elke vraag bijdraagt aan het meten van het gevonden domein, en dus niet overbodig is. Ook de inter-schaal correlaties, die lager dan 0.79 waren gaven aan dat de domeinen op zichzelf staande domeinen waren omdat deze minder dan 60% overlapten. De resultaten van de factoranalyses geven bewijs voor het onderdeel 'extrapoleren'. De positieve associatie tussen de numerieke scores die artsen verkregen en de geschreven feedback toonde aan dat artsen die hoge scores hadden gekregen, ook inderdaad veelal positief commentaar kregen. Bewijs voor het 'generaliseren' van de resultaten was gevonden door het uitvoeren van generaliseerbaarheid analyses. Met deze analyses bleek dat voor het genereren van een betrouwbare gemiddelde score voor artsen, beoordelingen van minimaal drie medisch specialist-collega's, twee AIOS en drie medewerkers nodig was.

In **hoofdstuk 4** is er verder onderzoek gedaan naar het bewijs van 'extrapoleren' voor het gebruik van vragenlijsten. In dit onderzoek is er gekeken naar een aspect van het onderdeel 'extrapoleren' wat nog niet onderzocht was. Het betreft hier de associatie tussen de 'subjectieve' MSF scores van artsen met 'objectieve' maatstaven vanuit de praktijk. Oftewel: krijgen artsen die goed functioneren op basis van klinische uitkomsten, ook hoge MSF scores van hun collega's? Om dit te onderzoeken is het klinisch functioneren en de beoordelingen van 28 anesthesiologen onderzocht. In 2014 hadden deze anesthesiologen 8030 anesthesie procedures uitgevoerd, waaruit het gemiddelde functioneren op basis van vijf kwaliteitsmaten kon worden berekend. Deze vijf klinische kwaliteitsmaten waren vooraf bepaald op basis van literatuur, onderzoek en protocollen en geven een indicatie van het perioperatieve functioneren van anesthesiologen. Het betreffen twee uitkomstmaten en drie procesmaten, namelijk (1) intraoperatieve pijn management, (2) preventie van postoperatieve misselijkheid en braken, (3) intraoperatieve temperatuur monitoring, (4) handhaving van de normale lichaamstemperatuur tijdens de operatie, en (5) de neuromusculaire functie monitoring. In datzelfde jaar zijn de 28 anesthesiologen door 56 AIOS, 38 anesthesiologen, 69 andere medisch specialisten en 144 medewerkers van multisource feedback voorzien, door middel van de 'INCEPT'. Ook hier zijn de drie domeinen van functioneren -patiëntgerichtheid, professionele houding, en (zelf)management en organisatorische vaardigheden- per type beoordelaar meegenomen in de analyses. De resultaten van dit onderzoek laten zien dat de relatie tussen 'subjectieve' maten en 'objectieve' maten complex is. Zo blijkt uit de multilevel regressie analyses dat de relatie tussen deze maten verschilt per type beoordelaar en per type domein van het functioneren. Zo

geven AIOS hogere MSF scores voor het domein professionele houding aan anesthesiologen die gemiddeld beter de normale lichaamstemperatuur van patiënten handhaafden, terwijl andere medewerkers juist lagere scores geven aan deze anesthesiologen. Ook krijgen anesthesiologen, die gemiddeld beter de temperatuur van patiënten onder narcose monitoren, hogere MSF scores voor hun (zelf)management en organisatorische vaardigheden van AIOS maar niet van hun collega's uit een ander specialisme. Over de patiëntgerichtheid van anesthesiologen zijn alle beoordelaars het wel eens: anesthesiologen die gemiddeld vaker de lichaamstemperatuur van patiënten onder narcose monitoren en vaker preventiemaatregelen uitvoeren om patiënten hun postoperatieve misselijkheid en braken te voorkomen, krijgen van alle type beoordelaars een hogere MSF score voor hun patiëntgerichtheid. Deze bevindingen tonen aan dat de associaties tussen 'subjectieve' MSF scores en 'objectieve' klinische maatstaven niet zo eenvoudig zijn. Elk type beoordelaar is het eens dat hoe beter anesthesiologen de temperatuur van patiënten onder narcose monitoren en preventiemaatregelen nemen om postoperatieve misselijkheid en braken te voorkomen, hoe hoger zij scoren op patiëntgerichtheid. Echter, voor de andere domeinen van functioneren zijn de associaties tussen de 'subjectieve' en 'objectieve' maten complexer en moet er rekening gehouden worden met welk perspectief de beoordelaar naar het functioneren van anesthesiologen kijkt.

De laatste stap in het onderzoek naar het validiteitsargument was het onderzoeken van het vierde en laatste onderdeel: de implicaties van het gebruik van MSF voor artsen. In essentie is het doel van MSF, wanneer het gebruikt wordt voor formatieve doeleinden, om artsen daar waar nodig hun functioneren te laten verbeteren op basis van de gekregen feedback. Met deze feedback van hun beoordelaars komen belangrijke tekortkomingen in het functioneren aan het licht, terwijl er tegelijk ook complimenten gegeven kunnen worden. Voor het onderdeel 'implicaties' moet er daarom bewijs worden gezocht over de gevolgen van het gebruik van vragenlijsten voor het geven van feedback, waar in **hoofdstuk 5** nader wordt ingegaan. Met een observationele studie is er onderzocht of het functioneren van artsen verbeterd, nadat deze artsen zijn beoordeeld met MSF en deze feedback naderhand hebben gekregen. In de periode van 2012 tot 2018 zijn 103 artsen tweemaal beoordeeld met MSF, in totaal door 242 AIOS, 684 collega medisch specialisten, en 999 medewerkers. Deze artsen hebben ook allen een zelfbeoordeling uitgevoerd, om hun eigen functioneren te beoordelen. In deze studie hebben we specifiek gekeken naar de mogelijke gevolgen van uiteenlopende feedback tussen deze zelf en anderen-beoordelingen, en dan met name wanneer artsen zichzelf hoger beoordeelden dan hun beoordelaars hen beoordeelden. Met MSF kunnen artsen worden geconfronteerd met feedback die niet strookt met hun eigen overtuigingen. Deze incongruentie kan zowel positief als negatief zijn, wat betekent dat artsen hun eigen prestaties respectievelijk onderschatten of overschatten. Vooral deze negatieve discrepanties tussen de zelf-scores en beoordelaars-scores zijn

interessant om in overweging te nemen als we kijken naar de gevolgen van MSF, omdat deze ofwel een positieve gedragsverandering kunnen stimuleren of destructief kunnen zijn voor toekomstig functioneren. Enerzijds kunnen negatieve discrepanties tussen de zelf-scores van artsen en de scores van de beoordelaars gunstig zijn voor artsen, aangezien ze onbekende tekortkomingen aan het licht brengen. Aan de andere kant kunnen artsen wanneer ze worden geconfronteerd met negatieve discrepanties, ook emotionele stress ervaren die juist ongunstig kan zijn voor het accepteren van de feedback, en zodoende lastig maakt om tot verbetering te komen. Tot op heden was er weinig bekend over de invloed van deze negatieve discrepanties op de professionele prestaties van artsen met betrekking tot MSF. Met behulp van multilevel analyses zijn de associaties tussen deze negatieve discrepanties en de verandering in daaropvolgende MSF-scores gekwantificeerd. Wederom is voor het verzamelen van MSF de INCEPT gebruikt, waarbij de gemiddelde score van artsen is onderverdeeld in drie domeinen -patiëntgerichtheid, professionele houding, en (zelf)management en organisatorische vaardigheden-. Zo is er onderzocht wat voor invloed het aantal negatieve discrepanties, waar artsen mee geconfronteerd worden tijden het krijgen van feedback, heeft op hun gemiddelde domein scores in de tweede MSF beoordelingsronde. Het aantal negatieve discrepanties is berekend door te tellen hoe vaak artsen zichzelf overschatten op de 18 stellingen waar artsen zelf en hun beoordelaars een score op moeten geven. Bij elke stelling kunnen artsen zichzelf overschatten per type beoordelaar, dus vergeleken met de scores verkregen van AIOS, collega medisch specialisten en medewerkers kunnen artsen zichzelf overschatten. In de analyses is er rekening gehouden met de invloed van de verschillende type beoordelaars. Uit de resultaten bleek dat het aantal negatieve discrepanties een significante negatieve relatie heeft met score veranderingen, in alle drie de professionele domeinen. Dit betekent dat wanneer artsen worden geconfronteerd met meerdere negatieve discrepanties, de mate van verbetering van artsen afneemt en bij een teveel aan negatieve discrepanties zelfs geen verbetering optreedt. Dit was het geval voor de scores van alle type beoordelaars. Artsen die zichzelf dus overschatten in de eerste beoordelingsronde vertonen in de tweede beoordelingsronde minder verbetering in hun functioneren, tegenover artsen die zichzelf niet hadden overschat. Een mogelijke verklaring voor dit gevonden resultaat kan zijn dat artsen die zichzelf overschatten (te)veel zelfvertrouwen hebben, wat het accepteren van incongruente feedback kan bemoeilijken. Uit eerder onderzoek is gebleken dat teveel zelfvertrouwen er voor kan zorgen dat de feedback, vooral wanneer deze incongruent is, wordt afgewezen en als 'onwaar' wordt bestempeld. De resultaten uit ons onderzoek vragen extra aandacht voor de follow-up van artsen na het verkrijgen van MSF, vooral voor artsen die zichzelf overschatten.

In het laatste hoofdstuk, **hoofdstuk 6**, zijn de resultaten van de voorgaande studies samengevat, geanalyseerd en gesynthetiseerd om een antwoord te geven op de onderzoeksvraag: "Welk bewijs moet er worden verzameld, ter ondersteuning of

weerlegging van het validiteitsargument voor het gebruik van vragenlijsten om artsen hun functioneren te evalueren en te beoordelen?". Het antwoord op deze vraag behoeft een analyse waarbij rekening gehouden moet worden met verschillende perspectieven op dit vraagstuk. Het post-positivistische perspectief ziet bewijs van geen meetfouten tijdens de beoordeling als sterk bewijs, terwijl dit voor het socio-constructivistische perspectief minder sterk wordt bezien: immers, het functioneren van artsen is niet in één ware score te vatten. Het antwoord op de onderzoeksvraag is dan ook niet zo eenvoudig en gemakkelijk samen te vatten. Hoewel beide onderzoeks-standpunten zich verschillend verhouden tot het validiteitsbewijs, kan uit beide standpunten worden geconcludeerd dat het validiteitsargument voor het gebruik van vragenlijsten, inclusief multisource feedback, nog niet sterk genoeg is om belangrijke beslissingen te nemen over artsen hun functioneren. Uiteraard moet het antwoord op de onderzoeksvraag en de generalisatie van de onderzoeksbevindingen gezien worden met in acht neming van de beperkingen in dit onderzoek. Om het gebruik van vragenlijsten, zowel voor het geven van feedback en het maken van beslissingen te bevorderen is er een alternatief model nodig voor beoordeling: het model van programmatisch toetsen. Programmatisch toetsen vraagt om verschillende beoordelingsmethodes die zorgvuldig zijn gecombineerd en geconstrueerd als een beoordelingsprogramma, bedoeld om het complete en complexe palet van het professionele functioneren van de arts vast te leggen. Hoe dit precies in de praktijk eruit ziet, zal vooraf goed worden onderzocht waarbij advies kan worden ingewonnen uit voorgaand onderzoek bij geneeskunde studenten. Hoofdstuk 6 eindigt met een gezegde: "Great minds think alike - but fools rarely differ". Hoewel dit gezegde eigenlijk aangeeft dat wanneer twee mensen hetzelfde idee hebben, ze ofwel briljant of dwaas kunnen zijn, wil ik ook graag zeggen dat briljante mensen misschien wel hetzelfde denken, maar dat alleen dwazen zelden een ander perspectief gebruiken.